The Annals of Statistics
1993, Vol. 21, No. 1, 355–377

INVARIANTS OF SOME PROBABILITY MODELS USED IN PHYLOGENETIC INFERENCE

BY STEVEN N. EVANS[1] AND T. P. SPEED[2]

University of California, Berkeley

The so-called method of invariants is a technique in the field of molecular evolution for inferring phylogenetic relations among a number of species on the basis of nucleotide sequence data. An invariant is a polynomial function of the probability distribution defined by a stochastic model for the observed nucleotide sequence. This function has the special property that it is identically zero for one possible phylogeny and typically nonzero for another possible phylogeny. Thus it is possible to discriminate statistically between two competing phylogenies using an estimate of the invariant. The advantage of this technique is that it enables such inferences to be made without the need for estimating nuisance parameters that are related to the specific mechanisms by which the molecular evolution occurs. For a wide class of models found in the literature, we present a simple algebraic formalism for recognising whether or not a function is an invariant and for generating all possible invariants. Our work is based on recognising an underlying group structure and using discrete Fourier analysis.

1. Introduction. The problem of inferring phylogenetic relations among a group of species using nucleotide sequence data is one of continuing interest to researchers in the field of molecular evolution. There are a variety of approaches to the problem in current use, see Swofford and Olsen (1990) for a recent review, and our concern is with methods based upon simple probabilistic models for nucleotide substitution. Such models have been in use for some time now, but interest in them heightened following the revelation by Felsenstein (1978) that the popular parsimony criterion can give rise to serious biases when the rates of evolutionary change in the true phylogenetic tree differ greatly from one branch to another.

In our view, the use of statistical models fitted by maximum likelihood is currently the best method of inferring phylogenies [see, e.g., Felsenstein (1981), Tavaré (1986), Barry and Hartigan (1987), and Navidi, Churchill and von Haeseler (1992)]. However, in recent years much interest has focussed on a simpler approach using functions of the data which permit inferences

Received April 1991; revised December 1991.
[1]Partially supported by NSF Grant DMS-90-15708.
[2]Partially supported by NSF Grant DMS-88-02378.
AMS 1991 *subject classifications*. Primary 62H05; secondary 60K99, 62F99.
Key words and phrases. Invariant, phylogenetic inference, discrete Fourier analysis, random walk on a group.

concerning the phylogeny without requiring the estimation of other parameters describing the nucleotide substitution mechanism. This approach has been called the *method of invariants* and we may describe it informally as follows (a full description is given in Section 2).

Suppose that we have aligned DNA sequence data for a number of taxa. For a given position in the sequence (typically, the third, second or first codon position of a DNA sequence coding for a common protein such as cytochrome c) we have a stochastic model for the observed base. This model is built using two ingredients. The first ingredient is a dependence structure reflecting the putative phylogeny. The second ingredient is a collection of stochastic mechanisms describing the occurrence of base substitution events along the branches of the phylogenetic tree. An *invariant* is a polynomial function that has as its argument the probability distribution of the observed bases and that, for a particular phylogeny, is zero for all choices of the substitution mechanisms. If it is assumed that the bases at different positions are i.i.d., then it is easy to estimate such an invariant without estimating the parameters describing the base substitution mechanism; and, if the invariant is typically nonzero for another specification of the phylogeny, then it is possible to discriminate statistically between the two competing phylogenies. Moreover, one of the hopes for the method of invariants is that the assumption of identical distribution for different sites can be weakened—a generalisation that does not seem as feasible with maximum likelihood methods.

Invariants were first defined by Cavender and Felsenstein (1987) and Lake (1987) for models involving four taxa. These and subsequent attempts at finding invariants have been, to a certain extent, ad hoc. In order to fully exploit the potential of the method of invariants it is necessary to have techniques for generating all possible invariants and for recognising when a given function is an invariant. The purpose of the present paper is to describe simple algebraic procedures that achieve these ends.

The outline of the remainder of the paper is as follows. In Section 2 we formally describe the models we will be dealing with and formally define what we mean by an invariant. Having developed the relevant nomenclature, we give a comparison of our work and previous work in the area. We also make the key observation that there is a group structure inherent in the models we are considering. With this in mind, we digress in Sections 3 and 4 to give a brief overview of discrete Fourier theory and random walks on finite groups, respectively. In Section 5 we give another description of the models in group language. In Section 6 we present our procedures for constructing and recognising invariants. We discuss examples involving two, three and four taxa in Section 7. A noteworthy feature of these examples is that the number of algebraically independent invariants always coincides with the number of "degrees of freedom" obtained by an informal parameter counting argument. (Some care needs to be taken when doing this counting due to issues of over-parametrisation and parameter identifiability.) We conjecture that the equality of these two numbers is a general phenomenon, but we do not as yet have a proof.

2. Definitions and notation. Suppose that we have aligned DNA sequence data for m taxa. We may construct a general class of stochastic models for the bases observed at a given position in the following manner.

Consider a finite rooted tree $\mathbf{T}$ with m leaves. Let $\mathbf{V}$ denote the set of vertices of $\mathbf{T}$. Write ρ for the root of $\mathbf{T}$ and $\mathbf{L}$ for the set of leaves of $\mathbf{T}$. For each vertex $v \in \mathbf{V} \setminus \{\rho\}$, there is a unique vertex $\sigma(v)$ which is connected to v by an edge and is closer to ρ in the usual graph-theoretic distance. Write $(\sigma(v),v)$ for the unique edge which connects $\sigma(v)$ and v.

Label the taxa with the elements of $\mathbf{L}$ and think of the collection of observed bases as a realisation of a $\{A,G,C,T\}^{\mathbf{L}}$-valued random variable $(Y_l)_{l \in \mathbf{L}}$ with a distribution defined as follows. Let π be a probability distribution on $\{A,G,C,T\}$. We will refer to π as the *root distribution*. For each vertex $v \in \mathbf{V} \setminus \{\rho\}$, let $P^{(v)}$ be a stochastic matrix on $\{A,G,C,T\}$. We will refer to $P^{(v)}$ as the *substitution matrix* associated with the edge $(\sigma(v), v)$. Define a probability distribution μ on $\{A,G,C,T\}^{\mathbf{V}}$ by setting

$$\mu\big((i_v)_{v \in \mathbf{V}}\big) = \pi(i_\rho) \prod_{v \in \mathbf{V} \setminus \{\rho\}} P^{(v)}\big(i_{\sigma(v)}, i_v\big).$$

Finally, let $(Y_l)_{l \in \mathbf{L}}$ have the marginal distribution

$$\mathbb{P}\big\{(Y_l)_{l \in \mathbf{L}} = (i_l)_{l \in \mathbf{L}}\big\} = \sum_{v \in \mathbf{V} \setminus \mathbf{L}} \sum_{i_v} \mu\big(\big((i_v)_{v \in \mathbf{V} \setminus \mathbf{L}}, (i_l)_{l \in \mathbf{L}}\big)\big),$$

where each of the dummy variables i_v, $v \in \mathbf{V} \setminus \mathbf{L}$, is summed over the set $\{A,G,C,T\}$.

The various elements appearing in these models have the following interpretations. The tree $\mathbf{T}$ is a candidate for the true phylogenetic tree describing the evolution of the observed present-day species corresponding to the leaves of the tree, insofar as this evolution is indicated by the evolution of the aligned sequence of nucleotides under study. The root of the tree ρ corresponds to an unobserved common ancestor of all of the observed present-day species, whilst the vertices other than the root and the leaves correspond to unobserved species intermediate in the evolutionary process, being common ancestors of pairs, triples, and so on of the observed present-day species. The root distribution π is thought of as the relative frequency of bases in the common ancestor's sequence, whilst the substitution matrices $P^{(v)}$ give a tractable and plausible probability model for the substitution process. We remark that the distribution μ on $\{A,G,C,T\}^{\mathbf{V}}$ satisfies a Markov property which may be stated as follows: for any two vertices v_1 and v_2, the base at v_1 and the base at v_2 are conditionally μ-independent given the base at any vertex v_3 on the unique path connecting v_1 and v_2.

The models of this form which appear in the literature usually take each substitution matrix to be the transition matrix at some point in time of a continuous time Markov chain on the state space $\{A,G,C,T\}$ (which particular point in time is possibly different for each edge, and these variables constitute "unknown parameters" in the model). Moreover, the Markov chain is usually taken to be from some subfamily of the possible chains on

$\{A, G, C, T\}$. The subfamilies we will be particularly interested in are described most easily in terms of the infinitesimal generator matrix of the chain. Kimura (1981) presents a model in which the infinitesimal generator matrix is of the form

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix} \end{array}$$

where $\alpha, \beta, \gamma \geq 0$. The value of the triple (α, β, γ) is possibly different for each edge, and these variables also constitute "unknown parameters" in the model. We will refer to this model as the *Kimura three-parameter model*. If we further restrict the class of allowable infinitesimal generator matrices by imposing the extra condition that $\beta = \gamma$, then we obtain the model considered by Kimura (1980). We will refer to this model as the *Kimura two-parameter model*. Finally, if we require that $\alpha = \beta = \gamma$ we obtain the model considered in Jukes and Cantor (1969) and more explicitly in Neyman (1971), which we will refer to as the *Jukes–Cantor model*.

As yet we have not said anything about the choice of the root distribution π. We will take π to be either the uniform distribution on $\{A, G, C, T\}$ or otherwise some arbitrary (and "unknown") probability distribution on $\{A, G, C, T\}$. Note that all the Markov chains described in the previous paragraph are reversible with the uniform distribution as the symmetrising stationary measure. We do not make explicit use of reversibility in this paper.

Let F be a polynomial in the dummy variables t_i, $i \in \{A, G, C, T\}^{\mathbf{L}}$. We say that F is an *invariant* for one of the models defined above if $F((\mathbb{P}\{Y = i\})_{i \in \{A,G,C,T\}^{\mathbf{L}}}) = 0$ for all choices of parameters in the model. We described the statistical uses of invariants in Section 1. The concept was introduced by Cavender and Felsenstein (1987) and Lake (1987). Cavender and Felsenstein (1987) and later Drolet and Sankoff (1990), Sankoff (1990) and Felsenstein (1991) derived invariants for Jukes–Cantor models with uniform root distribution and at most five taxa. Lake (1987) and later Cavender (1989, 1991) obtained linear invariants for a four taxa model based on a parametric family of substitution matrices that contains the Kimura two-parameter and Jukes–Cantor families. [We will show in Section 7 that, contrary to a claim made in Cavender (1991), there can be strictly fewer linear invariants for the Cavender–Lake model than there are for the Kimura two-parameter model.]

EXAMPLE. Consider the tree in Figure 1 with the leaves labelled as 1, 2, 3, 4 and the root labelled as 0. Suppose that we have a Jukes–Cantor model with uniform root distribution constructed from this tree. In general there are only 15 distinct probabilities $\mathbb{P}\{Y = (i_1, i_2, i_3, i_4)\}$, corresponding to the possible partitions of $\{1, 2, 3, 4\}$ defined by the equalities and inequalities amongst

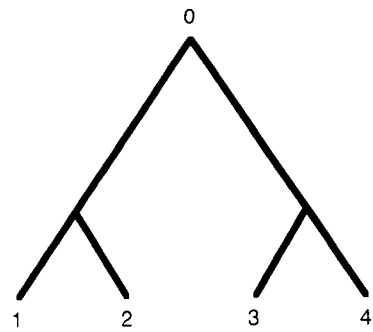

FIG. 1.

i_1, i_2, i_3, i_4. Write these 15 values as f_{1234}, $f_{1|234}$ and three similar, $f_{12|34}$ and two similar, $f_{12|3|4}$ and five similar, and $f_{1|2|3|4}$, where, for example, $f_{1234} = \mathbb{P}\{Y = (A, A, A, A)\} = \cdots = \mathbb{P}\{Y = (T, T, T, T)\}$. Lake (1987) shows that

$$f_{13|24} - f_{13|2|4} - f_{24|1|3} + f_{1|2|3|4} = 0$$

and

$$f_{14|23} - f_{14|2|3} - f_{1|4|23} + f_{1|2|3|4} = 0,$$

and these observations can be used to construct linear invariants for this model.

As we stated in Section 1, our aim in this paper is to describe a relatively simple algebraic formalism for generating/recognising all invariants for any of the models considered above when there is an arbitrary number of taxa. The key to our approach is the following observation. Suppose that we think of the bases $\{A, G, C, T\}$ as the elements of an Abelian (that is, commutative) group with the group operation defined by the following addition table:

$$\begin{array}{c|cccc} + & A & G & C & T \\ A & A & G & C & T \\ G & G & A & T & C \\ C & C & T & A & G \\ T & T & C & G & A \end{array}.$$

This group is isomorphic to the *Klein four-group* $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ [i.e., the group consisting of the elements $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ with the group operation being coordinatewise addition modulo 2]. One possible isomorphism is given by $A \leftrightarrow (0, 0)$, $G \leftrightarrow (0, 1)$, $C \leftrightarrow (1, 0)$ and $T \leftrightarrow (1, 1)$. It is straightforward to check that the infinitesimal generator matrices appearing in the Kimura three-parameter model have the property that the entry corresponding to the pair of

bases (i, j) is a function of the base $i - j$. The same is also true a fortiori for the Kimura two-parameter model and the Jukes–Cantor model. Thus such a matrix is nothing other than the infinitesimal generator matrix for a random walk on the group (see Section 4 for a discussion of continuous time random walks on finite Abelian groups).

3. Some discrete Fourier analysis. For the sake of reference, we review some elementary facts regarding Fourier analysis on finite Abelian groups. There seems to be no good reference to Fourier analysis and group characters which treats the Abelian case in isolation and contains all the theory we need. However, Chapter 104 of Körner (1988) is a good introduction, and all of what we have to say here may be deduced fairly easily from the more general material presented in Chapter 2 of Diaconis (1988). In the same spirit as our work, Diaconis (1990) investigates how general Fourier theory may be used to analyse the properties of patterned matrices when the pattern reflects invariance under the action of some group.

Let $\mathbb{G}$ be a finite Abelian group, with the group operation written as $+$. Let $\mathbb{T} = \{z \in \mathbb{C}: |z| = 1\}$ denote the unit circle in the complex plane, and regard $\mathbb{T}$ as an Abelian group with the group operation being ordinary complex multiplication. The *characters* of $\mathbb{G}$ are the group homomorphisms mapping $\mathbb{G}$ into $\mathbb{T}$. That is, $\chi: \mathbb{G} \to \mathbb{T}$ is a character if $\chi(g_1 + g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in \mathbb{G}$. The characters form an Abelian group under the operation of pointwise multiplication of functions. This group is called the *dual group* of $\mathbb{G}$ and is denoted by $\hat{\mathbb{G}}$. The groups $\mathbb{G}$ and $\hat{\mathbb{G}}$ are isomorphic. Given $g \in \mathbb{G}$ and $\chi \in \hat{\mathbb{G}}$, write $\langle g, \chi \rangle$ for $\chi(g)$. The dual of the direct sum $\mathbb{G}^m = \oplus_{i=1}^m \mathbb{G}$ is isomorphic to $\hat{\mathbb{G}}^m$ under the isomorphism given by $\langle (g_1, \ldots, g_m), (\chi_1, \ldots, \chi_m) \rangle = \prod_{i=1}^m \langle g_i, \chi_i \rangle$.

EXAMPLE. Suppose that $\mathbb{G} = \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Then one may write $\hat{\mathbb{G}} = \{1, \phi, \psi, \phi\psi\}$, where the following table gives the value of $\langle g, \chi \rangle$ for $g \in \mathbb{G}$ and $\chi \in \hat{\mathbb{G}}$:

$$\begin{array}{c} \\ 1 \\ \phi \\ \psi \\ \phi\psi \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \end{array}.$$

Given a function $f: \mathbb{G} \to \mathbb{C}$, the function $\hat{f}: \hat{\mathbb{G}} \to \mathbb{C}$ defined by

$$\hat{f}(\chi) = \sum_{g \in \mathbb{G}} \langle g, \chi \rangle f(g)$$

is called the *Fourier transform* of f. If f is a discrete probability mass function on $\mathbb{G}$ and Z is a $\mathbb{G}$-valued random variable with distribution f, then

$\hat{f}(\chi) = \mathbb{E}[\langle Z, \chi\rangle]$. The Fourier transform has the following properties:

$$\hat{1}(\chi) = \begin{cases} |\mathbb{G}|, & \text{if } \chi = 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $|A|$ denotes the cardinality of a set A; and

$$\widehat{f_1 * f_2} = \hat{f}_1 \hat{f}_2,$$

where

$$f_1 * f_2(g) = \sum_{h \in \mathbb{G}} f_1(g - h) f_2(h), \qquad g \in \mathbb{G},$$

is the convolution of the functions f_1 and f_2. Moreover, a function may be recovered from its Fourier transform by the process of *Fourier inversion*; namely, if f has Fourier transform $\hat{f}$ then $f(g) = |\mathbb{G}|^{-1} \Sigma_{\chi \in \hat{\mathbb{G}}} \overline{\langle g, \chi\rangle} \hat{f}(\chi)$ for all $g \in \mathbb{G}$.

4. Random walks. Suppose that $\mathbb{G}$ is a finite Abelian group and $X = (X_t, \mathbb{P}^g)$ is a continuous time Markov chain on $\mathbb{G}$ (here, $\mathbb{P}^g$, $g \in \mathbb{G}$, is the probability measure on path-space corresponding to starting the chain off at the initial point g). Let P_t denote the corresponding semigroup of transition matrices (i.e., $P_t(i, j) = \mathbb{P}^i\{X_t = j\}$) and let Q denote the corresponding infinitesimal generator matrix. We say that the process X is a *random walk* if, for all $t \geq 0$ and all $i, j \in \mathbb{G}$, $P_t(i, j) = p_t(j - i)$ for some probability distribution p_t on $\mathbb{G}$ or, equivalently, that $Q(i, j) = q(i - j)$ for some function q: $\mathbb{G} \to \mathbb{R}$ such that $\Sigma_g q(g) = 0$.

We can describe such a process probabilistically as follows. Let N be a simple, homogeneous Poisson process with rate $-q(0)$ and let $\{J_n\}_{n=1}^{\infty}$ be an independent sequence of i.i.d. $\mathbb{G}$-valued random variables with common distribution given by

$$\mathbb{P}\{J_n = j\} = \begin{cases} q(j) \Big/ \Big(\sum_{g \neq 0} q(g) \Big), & \text{if } j \neq 0, \\ 0, & \text{if } j = 0. \end{cases}$$

Then the distribution of $\{X_t\colon t \geq 0\}$ under $\mathbb{P}^g$ is the same as the distribution of $\{g + \Sigma_{n=1}^{N_t} J_n\colon t \geq 0\}$, where we define the empty sum to be 0. More generally, the distribution of $\{X_t\colon t \geq 0\}$ under $\mathbb{P}^\nu$, where ν is some arbitrary initial distribution, is the same as the distribution of $\{J_0 + \Sigma_{n=1}^{N_t} J_n\colon t \geq 0\}$, where J_0 is a $\mathbb{G}$-valued random variable with distribution ν and J_0 is independent of N and $\{J_n\}_{n=1}^{\infty}$.

From this description of X it is easy to see that we have the *Lévy–Hinčin formula*

$$\hat{p}_t(\chi) = \exp\Big(t \sum_{g \in \mathbb{G} \setminus 0} (\langle g, \chi\rangle - 1) q(g) \Big) = \exp\Big(t \sum_{g \in \mathbb{G}} \langle g, \chi\rangle q(g) \Big).$$

We are particularly interested in the case when $\mathbb{G}$ is the Klein four-group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. Here the matrix Q will be of the form

$$Q = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix} \end{array},$$

for some parameters $\alpha, \beta, \gamma \geq 0$. If we label the characters of $\mathbb{G}$ in the same way as we did in Section 3, then we see from the Lévy–Hinčin formula that

$$\begin{aligned} \hat{p}_t(1) &= 1, \\ \hat{p}_t(\phi) &= \exp(-2t(\alpha+\gamma)), \\ \hat{p}_t(\psi) &= \exp(-2t(\beta+\gamma)) \end{aligned}$$

and

$$\hat{p}_t(\phi\psi) = \exp(-2t(\alpha+\beta)).$$

Applying Fourier inversion, we see that

$$\begin{aligned} p_t((0,0)) &= \tfrac{1}{4}[1 + \exp(-2t(\alpha+\gamma)) \\ &\qquad + \exp(-2t(\beta+\gamma)) + \exp(-2t(\alpha+\beta))], \\ p_t((0,1)) &= \tfrac{1}{4}[1 - \exp(-2t(\alpha+\gamma)) \\ &\qquad + \exp(-2t(\beta+\gamma)) - \exp(-2t(\alpha+\beta))], \\ p_t((1,0)) &= \tfrac{1}{4}[1 + \exp(-2t(\alpha+\gamma)) \\ &\qquad - \exp(-2t(\beta+\gamma)) - \exp(-2t(\alpha+\beta))] \end{aligned}$$

and

$$\begin{aligned} p_t((1,1)) &= \tfrac{1}{4}[1 - \exp(-2t(\alpha+\gamma)) \\ &\qquad - \exp(-2t(\beta+\gamma)) + \exp(-2t(\alpha+\beta))]. \end{aligned}$$

Define R_3 to be the set of all the probability distributions on $\mathbb{G}$ which can occur as the distributions p_t if we let α, β, γ and t range over all possible values. Define R_2 (resp., R_1) similarly, but with the restriction that $\beta = \gamma$ (resp., $\alpha = \beta = \gamma$). The following lemmas are trivial given the above calculations, but they will be crucial ingredients in our procedure for constructing all possible invariants.

LEMMA 4.1. *The set* $\{(\hat{r}(\phi), \hat{r}(\psi), \hat{r}(\phi\psi))\colon r \in R_3\} \subset \mathbb{R}^3$ *has a nonempty interior.*

LEMMA 4.2. *The equality* $\hat{r}(\phi) = \hat{r}(\phi\psi)$ *holds for all* $r \in R_2$ *and the set* $\{(\hat{r}(\phi), \hat{r}(\psi))\colon r \in R_2\} \subset \mathbb{R}^2$ *has a nonempty interior.*

LEMMA 4.3. *The equality* $\hat{r}(\phi) = \hat{r}(\psi) = \hat{r}(\phi\psi)$ *holds for all* $r \in R_1$ *and the set* $\{\hat{r}(\phi)\colon r \in R_1\} \subset \mathbb{R}$ *has a nonempty interior.*

5. Another description of the models. Identify the four bases $\{A, G, C, T\}$ with the elements of the Klein four-group $\mathbb{G} = \mathbb{Z}_2 \oplus \mathbb{Z}_2$ as we did in Section 2. Each substitution matrix appearing in the description of the Kimura three-parameter model is thus of the form $P^{(v)}(i, j) = r^{(v)}(j - i)$ for some probability distribution $r^{(v)} \in R_3$. The same is true of the Kimura two-parameter model and the Jukes–Cantor model if we replace R_3 by R_2 and R_1, respectively.

Construct independent $\mathbb{G}$-valued random variables $(Z_v)_{v \in \mathbf{V}}$ such that Z_v has the distribution $r^{(v)}$ for each $v \in \mathbf{V} \setminus \{\rho\}$, and Z_ρ has the root distribution π. For each vertex $v \in \mathbf{V}$, let $\delta(v)$ denote the sequence of states along the unique path through the tree connecting ρ and v [we include both ρ and v in $\delta(v)$]. Then it is clear that the probability distribution μ from Section 1 is the distribution of the random variables $(\Sigma_{u \in \delta(v)} Z_u)_{v \in \mathbf{V}}$, and hence the random variables $(Y_l)_{l \in \mathbf{L}}$ have the same distribution as $(\Sigma_{u \in \delta(l)} Z_u)_{l \in \mathbf{L}}$. In the future we will suppose that the random variables $(Y_l)_{l \in \mathbf{L}}$ have actually been constructed as $(\Sigma_{u \in \delta(l)} Z_u)_{l \in \mathbf{L}}$. Thus, if we set $Y = (Y_l)_{l \in \mathbf{L}}$ and $Z = (Z_v)_{v \in \mathbf{V}}$, then we have an "additive random effects model" $Y = DZ$, where D is an appropriate "design" matrix of 0's and 1's. Here, of course, we are using the usual $\mathbb{Z}$-module notation $kg = \Sigma_{i=1}^{k} g$ for $k \in \mathbb{Z}$, $k \geq 0$ and $g \in \mathbb{G}$.

EXAMPLE. Suppose that $m = 4$ and $\mathbf{T}$ is the tree in Figure 2 with the vertices labelled as shown. If we take the vertex 0 as the root then the design matrix will be

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

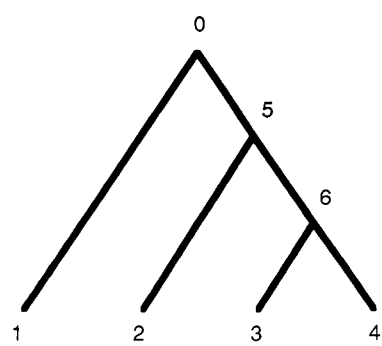

FIG. 2.

6. Constructing and classifying invariants. Let us begin by considering the Kimura three-parameter model with the uniform root distribution. Before presenting our general method for constructing invariants, we show how it works in a simple example. Suppose that $m = 3$ and $\mathbf{T}$ is the tree in Figure 3 with the vertices labelled as shown.

If we take the vertex 0 as the root then the model is

$$\begin{aligned} Y_1 &= Z_0 + Z_1, \\ Y_2 &= Z_0 + Z_2, \\ Y_3 &= Z_0 + Z_3, \end{aligned} \tag{6.1}$$

where $(Z_i)_{i=0}^3$ are independent, Z_0 has the uniform distribution on $\mathbb{G}$ and each Z_i, $i \in \{1, 2, 3\}$, has a distribution belonging to R_3.

Observe that

$$\begin{aligned} &\mathbb{E}\big[\langle Y_1, \phi\rangle\langle Y_2, \psi\rangle\langle Y_3, \phi\psi\rangle\big] \\ &\quad = \mathbb{E}\big[\langle Z_0 + Z_1, \phi\rangle\langle Z_0 + Z_2, \psi\rangle\langle Z_0 + Z_3, \phi\psi\rangle\big] \\ &\quad = \mathbb{E}\big[\langle Z_0, \phi\rangle\langle Z_0, \psi\rangle\langle Z_0, \phi\psi\rangle\langle Z_1, \phi\rangle\langle Z_2, \psi\rangle\langle Z_3, \phi\psi\rangle\big] \\ &\quad = \mathbb{E}\big[\langle Z_0 + Z_0, \phi\psi\rangle\langle Z_1, \phi\rangle\langle Z_2, \psi\rangle\langle Z_3, \phi\psi\rangle\big] \\ &\quad = \mathbb{E}\big[\langle Z_1, \phi\rangle\big]\mathbb{E}\big[\langle Z_2, \psi\rangle\big]\mathbb{E}\big[\langle Z_3, \phi\psi\rangle\big], \end{aligned}$$

where the second and third lines follow from the fact that the characters are homomorphisms and the last line follows from this fact, the fact that each element of $\mathbb{G}$ is its own inverse and the independence of $(Z_i)_{i=1}^3$. Similarly,

$$\mathbb{E}\big[\langle Y_1, \phi\psi\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle\big] = \mathbb{E}\big[\langle Z_1, \phi\psi\rangle\big]\mathbb{E}\big[\langle Z_2, \phi\rangle\big]\mathbb{E}\big[\langle Z_3, \psi\rangle\big]$$

and

$$\mathbb{E}\big[\langle Y_1, \psi\rangle\langle Y_2, \phi\psi\rangle\langle Y_3, \phi\rangle\big] = \mathbb{E}\big[\langle Z_1, \psi\rangle\big]\mathbb{E}\big[\langle Z_2, \phi\psi\rangle\big]\mathbb{E}\big[\langle Z_3, \phi\rangle\big].$$

Also observe, by similar reasoning, that

$$\mathbb{E}\big[\langle Y_i, \theta\rangle\langle Y_j, \theta\rangle\big] = \mathbb{E}\big[\langle Z_i, \theta\rangle\big]\mathbb{E}\big[\langle Z_j, \theta\rangle\big]$$

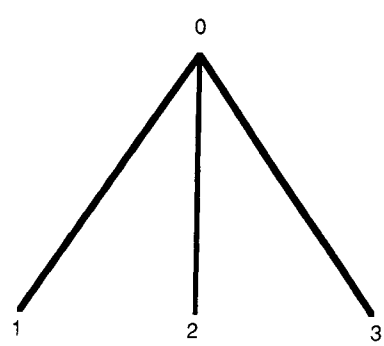

FIG. 3.

for $1 \le i < j \le 3$ and $\theta \in \{\phi, \psi, \phi\psi\}$. Therefore

$$\begin{aligned}\big(\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \psi\rangle\langle Y_3, \phi\psi\rangle]\mathbb{E}[\langle Y_1, \phi\psi\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle] \\ \times \mathbb{E}[\langle Y_1, \psi\rangle\langle Y_2, \phi\psi\rangle\langle Y_3, \phi\rangle]\big)^2 \\ - \prod_{1\le i<j\le 3}\ \prod_{\theta\in\{\phi,\psi,\phi\psi\}} \mathbb{E}\big[\langle Y_i, \theta\rangle\langle Y_j, \theta\rangle\big] = 0.\end{aligned}$$

Thus, if we express each of the expectations appearing above in terms of the variables $\mathbb{P}\{Y = g\}$, $g \in \mathbb{G}^3$, we see that we can construct a ninth degree polynomial F in the dummy variables $(t_g)_{g\in\mathbb{G}^3}$ such that $F[(\mathbb{P}\{Y = g\})_{g\in\mathbb{G}^3}]$ is identically zero for all possible choices of parameters in the model and hence we have found an invariant.

The rationale behind what we did here is to take two expressions of the form

$$\prod_{\chi}(\mathbb{E}[\langle Y, \chi\rangle])^{k(\chi;j)},$$

for $j = 1, 2$, and show that they will be equal for all possible choices of parameters in the model by showing that they may be reexpressed as a common product of powers of the quantities $\{\mathbb{E}[\langle Z_i, \theta\rangle]\colon\ 0 \le i \le 3,\ \theta \in \{\phi, \psi, \phi\psi\}\}$. Nowhere in constructing this invariant did we make use of the fact that π, the distribution of Z_0, is uniform on $\mathbb{G}$. In fact, no term of the form $(\mathbb{E}[\langle Z_0, \theta\rangle])^l$, $l \ge 1$, appeared in the common product.

Suppose now that if we express some multinomial $\Pi_\chi(\mathbb{E}[\langle Y, \chi\rangle])^{k(\chi)}$ as a product of powers of the quantities $\{\mathbb{E}[\langle Z_i, \theta\rangle]\colon 0 \le i \le 3, \theta \in \{\phi, \psi, \phi\psi\}\}$, then a term of the form $(\mathbb{E}[\langle Z_0, \theta\rangle])^l$, $l \ge 1$, does appear. We know from Section 3 that $\mathbb{E}[\langle Z_0, \theta\rangle] = 0$ for $\theta \in \{\phi, \psi, \phi\psi\}$. Thus $\Pi_\chi(\mathbb{E}[\langle Y, \chi\rangle])^{k(\chi)}$ is identically zero for all choices of parameters in the model and we have an invariant. For example, we have

$$\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \phi\rangle\langle Y_3, \phi\rangle] = \mathbb{E}[\langle Z_0, \phi\rangle]\mathbb{E}[\langle Z_1, \phi\rangle]\mathbb{E}[\langle Z_2, \phi\rangle]\mathbb{E}[\langle Z_3, \phi\rangle] = 0.$$

We will show below in Theorem 6.1 that these two ways of constructing invariants lead to all possible invariants for a general Kimura three-parameter model with uniform root distribution. First, however, we need some notation.

Given a character $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$, we can represent the function $z \mapsto \langle Dz, \chi\rangle$, $z \in \mathbb{G}^{\mathbf{V}}$, as $z \mapsto \Pi_{v\in\mathbf{V}}\langle z_v, \eta(v, \chi)\rangle$, where the characters $\eta(v, \chi) \in \hat{\mathbb{G}}$, $v \in \mathbf{V}$, are defined by $\eta(v, \chi) = \Pi_{l\in\mathbb{L}}\chi_l^{D_{l,v}}$. Moreover, this is the unique such representation. Let $\mathbf{H}$ denote the collection of multinomials in the dummy variables u_χ, $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$. That is, $h \in \mathbf{H}$ is of the form $h((u_\chi)) = \Pi_\chi u_\chi^{k_\chi}$, where $k_\chi \in \{0, 1, 2, \ldots\}$, $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$. Given such a multinomial $h \in \mathbf{H}$, we can uniquely define another multinomial S_3h in the dummy variables $w_{v,\theta}$, $v \in \mathbf{V}$ and $\theta \in \{\phi, \psi, \phi\psi\}$, by

$$S_3h((w_{v,\theta})) = \prod_{\chi}\Big(\prod_{v} W_3(v, \chi)\Big)^{k_\chi},$$

where we set

$$W_3(v,\chi) = \begin{cases} w_{v,\eta(v,\chi)}, & \text{if } \eta(v,\chi) \neq 1, \\ 1, & \text{if } \eta(v,\chi) = 1. \end{cases}$$

Observe that

$$h\big((\mathbb{E}[\langle Y,\chi\rangle])_\chi\big) = S_3 h\big((\mathbb{E}[\langle Z_v,\theta\rangle])_{v,\theta}\big).$$

We now define an equivalence relation $\sim_3$ on $\mathbf{H}$ as follows. Suppose that we have two multinomials $f, g \in \mathbf{H}$ with

$$S_3 f((w_{v,\theta})) = \prod_{v,\theta} w_{v,\theta^{a(v,\theta)}}$$

and

$$S_3 g((w_{v,\theta})) = \prod_{v,\theta} w_{v,\theta^{b(v,\theta)}}.$$

We declare that $f \sim_3 g$ if *either* $\Sigma_\theta a(\rho,\theta) \neq 0$ and $\Sigma_\theta b(\rho,\theta) \neq 0$, *or* $\Sigma_\theta a(\rho,\theta) = \Sigma_\theta b(\rho,\theta) = 0$ and $a(v,\theta) = b(v,\theta)$ for all $v \in \mathbf{V}\setminus\{\rho\}$ and all $\theta \in \{\phi,\psi,\phi\psi\}$. Write $\mathscr{H}_3$ for the family of equivalence classes in $\mathbf{H}$ under $\sim_3$, and let $\mathbf{H}_{3,\rho}$ denote the equivalence class consisting of multinomials h such that $S_3 h((w_{v,\theta}))$ is divisible by $w_{\rho,\xi}$ for some $\xi \in \{\phi,\psi,\phi\psi\}$. Observe that if $f, g \in \mathbf{H}$ with $f \sim_3 g$ then

$$f((\mathbb{E}[\langle Y,\chi\rangle])_\chi) = g((\mathbb{E}[\langle Y,\chi\rangle])_\chi),$$

for all choices of parameters in the model, with the common value being identically zero if $f, g \in \mathbf{H}_{3,\rho}$. Moreover, from Lemma 4.1 we see that the converse to this statement also holds.

Using this notation, we can describe the structure of the most general invariant as follows.

THEOREM 6.1. *Consider a Kimura three-parameter model with uniform root distribution. A polynomial F in the dummy variables $(t_g)_{g\in\mathbb{G}^{\mathbf{L}}}$ will be such that $F((\mathbb{P}\{Y=g\})_{g\in\mathbb{G}^{\mathbf{L}}}) = 0$ for all choices of parameters in the model if and only if F is of the form*

$$F((t_g)) = \sum_{h\in\mathbf{H}} c_h h\left(\left(\sum_{g\in\mathbb{G}^{\mathbf{L}}} \langle g,\chi\rangle t_g\right)_{\chi\in\hat{\mathbb{G}}^{\mathbf{L}}}\right),$$

where only finitely many of the coefficients c_h are non-zero and $\Sigma_{h\in\mathbf{K}} c_h = 0$ for all $\mathbf{K} \in \mathscr{H}_3 \setminus \{\mathbf{H}_{3,\rho}\}$.

PROOF. The sufficiency of the stated condition is already obvious from the observations we have made above.

Consider the question of necessity. Using Fourier inversion, we can express each variable t_g as a linear combination of the terms $\Sigma_{g\in\mathbb{G}^{\mathbf{L}}}\langle g,\chi\rangle t_g$, $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$,

and so we can certainly write

$$F\big((t_g)\big) = \sum_{h \in \mathbf{H}} c_h h\left(\left(\sum_{g \in \mathbb{G}^{\mathbf{L}}} \langle g, \chi \rangle t_g\right)_{\chi \in \hat{\mathbb{G}}^{\mathbf{L}}}\right)$$

for some coefficients (c_h), where only finitely many of the c_h are nonzero. For each $\mathbf{K} \in \mathscr{H}_3 \setminus \{\mathbf{H}_{3,\rho}\}$ let $k_{\mathbf{K}}$ denote the common value of $S_3 h$ for $h \in \mathbf{K}$. Recall that $w_{\rho,\xi}$ does not divide $k_{\mathbf{K}}((w_{v,\theta}))$ for any $\xi \in \{\phi, \psi, \phi\psi\}$; and so, from Lemma 4.1, the collection of functions on the space of parameters given by $k_{\mathbf{K}}((\mathbb{E}[\langle Z_v, \theta \rangle]))$, $\mathbf{K} \in \mathscr{H}_3 \setminus \{\mathbf{H}_{3,\rho}\}$, is linearly independent. As

$$\begin{aligned} F\big((\mathbb{P}\{Y = g\})\big) &= \sum_{h \in \mathbf{H}} c_h h\big((\mathbb{E}[\langle Y, \chi \rangle])\big) \\ &= \sum_{\mathbf{K} \in \mathscr{H}_3 \setminus \{\mathbf{H}_{3,\rho}\}} \left(\sum_{h \in \mathbf{K}} c_h\right) k_{\mathbf{K}}\big((\mathbb{E}[\langle Z_v, \theta \rangle])\big), \end{aligned}$$

the result follows. □

Given Theorem 6.1, we see that the problem of generating all invariants for the Kimura three-parameter model with uniform root distribution reduces to the two problems of:

(i) describing all multinomials $h \in \mathbf{H}_{3,\rho}$, and

(ii) describing all pairs of multinomials h' and h'' such that $h' \notin \mathbf{H}_{3,\rho}$, $h'' \notin \mathbf{H}_{3,\rho}$ and $h' \sim_3 h''$.

Regarding problem (i), observe that h belongs to $\mathbf{H}_{3,\rho}$ if and only if $h((u_\chi))$ is divisible by some u_{χ^*}, $\chi^* \in \hat{\mathbb{G}}^{\mathbf{L}}$, such that $\eta(\rho, \chi^*) \in \{\phi, \psi, \phi\psi\}$. Thus problem (i) reduces to computing and inspecting $\eta(\rho, \chi)$ for each $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$.

Problem (ii) is a little more involved. Let $\chi^{(1)}, \ldots, \chi^{(M)}$ be a list of the characters $\chi \in \hat{\mathbb{G}}^{\mathbf{L}}$ such that $\eta(\rho, \chi) = 1$. We can write two multinomials $h' \notin \mathbf{H}_{3,\rho}$ and $h'' \notin \mathbf{H}_{3,\rho}$ as

$$h'\big((u_\chi)\big) = \prod_{i=1}^{M} u_{\chi^{(i)}}^{a(i)}$$

and

$$h''\big((u_\chi)\big) = \prod_{i=1}^{M} u_{\chi^{(i)}}^{b(i)}.$$

Associate each character $\chi^{(i)}$ with a vector $x^{(i)}$ of 0's and 1's indexed by $(\mathbf{V} \setminus \{\rho\}) \times \{\phi, \psi, \phi\psi\}$ by setting

$$x_{v,\theta}^{(i)} = \begin{cases} 1, & \text{if } \eta\big(v, \chi^{(i)}\big) = \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently, $x_{v,\theta}^{(i)} = 1$ if and only if $W_3(v, \chi^{(i)}) = w_{v,\theta}$. Then $h' \sim_3 h''$ if and only if $\sum_{i=1}^{M} a(i) x^{(i)} = \sum_{i=1}^{M} b(i) x^{(i)}$. Observe that h' and h'' will be equivalent under $\sim_3$ if and only if the two multinomials obtained by removing common

factors are also equivalent; so, without loss of generality, it suffices to describe all pairs of nonnegative integer M-tuples $a = (a(i))_{i=1}^{M}$ and $b = (b(i))_{i=1}^{M}$ such that $\sum_{i=1}^{M} a(i)x^{(i)} = \sum_{i=1}^{M} b(i)x^{(i)}$ and at most one of the integers $a(i)$ and $b(i)$ is nonzero for each $i \in \{1, \ldots, M\}$. This latter problem is equivalent to describing the set

$$E_3 = \left\{ e = (e(i))_{i=1}^{M} \in \mathbb{Z}^M : \sum_{i=1}^{M} e(i)x^{(i)} = 0 \right\}.$$

It is clear that E_3 is a *lattice* in the sense of Section IV.3 in Cohn (1980). Therefore, by Theorem IV.5.1 of Cohn (1980), there exists a *minimal basis* for E_3 when $E_3 \neq \{0\}$. That is, there exist vectors $e^{(1)}, \ldots, e^{(N)} \in E_3$ such that E_3 consists of all vectors of the form $e = \sum_{j=1}^{N} n(j)e^{(j)}$ for some $(n(j))_{j=1}^{N} \in \mathbb{Z}^N$; and, moreover, the coefficients appearing in this representation are unique. The number N is intrinsic to E_3, that is, it is the same for all possible minimal bases. Here N coincides with the dimension of the real vector space

$$\left\{ f = (f(i))_{i=1}^{M} \in \mathbb{R}^M : \sum_{i=1}^{M} f(i)x^{(i)} = 0 \right\}.$$

Thus, $N = M - \text{rank}(C)$, where C is the matrix with ith row $x^{(i)}$. Cohn (1980) describes a procedure for constructing such a minimal basis.

The analogues of all of the above for the Kimura two-parameter model with uniform root distribution and the Jukes–Cantor model with uniform root distribution follow along similar lines, with Lemma 4.1 replaced by Lemma 4.2 and Lemma 4.3, respectively. We give a brief sketch of the main ideas and leave the details to the reader.

Given a multinomial $h \in \mathbf{H}$ of the form $h((u_\chi)) = \prod_\chi u_\chi^{k_\chi}$, where $k_\chi \in \{0, 1, 2 \ldots\}$, we can uniquely define another multinomial $S_2 h$ (resp., $S_1 h$) in the dummy variables $w_{v,\theta}$, $v \in \mathbf{V}$ and $\theta \in \{\phi, \psi\}$, (resp., in the dummy variables w_v, $v \in \mathbf{V}$) by $S_2 h((w_{v,\theta})) = \prod_\chi (\prod_v W_2(v, \chi))^{k_\chi}$ (resp., $S_1 h((w_v)) = \prod_\chi (\prod_v W_1(v, \chi))^{k_\chi}$), where we set

$$W_2(v, \chi) = \begin{cases} w_{v,\phi}, & \text{if } \eta(v, \chi) \in \{\phi, \phi\psi\}, \\ w_{v,\psi}, & \text{if } \eta(v, \chi) = \psi, \\ 1, & \text{if } \eta(v, \chi) = 1 \end{cases}$$

(resp.,

$$W_1(v, \chi) = \begin{cases} w_v, & \text{if } \eta(v, \chi) \neq 1, \\ 1, & \text{if } \eta(v, \chi) = 1). \end{cases}$$

Observe that for the Kimura two-parameter model

$$h\left((\mathbb{E}[\langle Y, \chi \rangle])_\chi\right) = S_2 h\left((\mathbb{E}[\langle Z_v, \theta \rangle])_{v,\theta}\right)$$

and that for the Jukes–Cantor model

$$h\left((\mathbb{E}[\langle Y, \chi \rangle])_\chi\right) = S_1 h\left((\mathbb{E}[\langle Z_v, \phi \rangle])_v\right).$$

We define equivalence relations $\sim_2$ and $\sim_1$ on $\mathbf{H}$ in the same manner as we defined $\sim_3$, but with S_3 replaced by S_2 and S_1, respectively.

Write $\mathscr{H}_2$ (resp., $\mathscr{H}_1$) for the family of equivalence classes in $\mathbf{H}$ under $\sim_2$ (resp., $\sim_1$). Let $\mathbf{H}_{2,\rho}$ denote the equivalence class in $\mathscr{H}_2$ consisting of multinomials h such that $S_2h((w_{v,\theta}))$ is divisible by $w_{\rho,\xi}$ for some $\xi \in \{\phi, \psi\}$ and let $\mathbf{H}_{1,\rho}$ denote the equivalence class in $\mathscr{H}_1$ consisting of multinomials h such that $S_1h((w_v))$ is divisible by w_ρ. Observe for the Kimura two-parameter model with uniform root distribution that if $f, g \in \mathbf{H}$ with $f \sim_2 g$ then

$$f((\mathbb{E}[\langle Y, \chi\rangle])_\chi) = g((\mathbb{E}[\langle Y, \chi\rangle])_\chi),$$

for all choices of parameters in the model, with the common value being identically zero if $f, g \in \mathbf{H}_{2,\rho}$. Similarly, for the Jukes–Cantor model with uniform root distribution observe that if $f, g \in \mathbf{H}$ with $f \sim_1 g$ then

$$f((\mathbb{E}[\langle Y, \chi\rangle])_\chi) = g((\mathbb{E}[\langle Y, \chi\rangle])_\chi),$$

for all choices of parameters in the model, with the common value being identically zero if $f, g \in \mathbf{H}_{1,\rho}$. Moreover, from Lemma 4.2 and Lemma 4.3 we see that the converses to the last two statements also hold.

For example, consider the Jukes–Cantor model with uniform root distribution which has the design specified by (6.1). It is easy to check that

$$\left(\mathbb{E}\left[\langle Y_1, \phi\rangle\langle Y_2, \psi\rangle\langle Y_3, \phi\psi\rangle\right]\right)^2 - \prod_{1\le i<j\le 3} \mathbb{E}\left[\langle Y_i, \phi\rangle\langle Y_j, \phi\rangle\right] = 0,$$

for all choices of parameters in the model.

The analogues of Theorem 6.1 for the Kimura two-parameter model and the Jukes–Cantor model are combined in the following result.

THEOREM 6.2. *Consider a Kimura two-parameter model (resp., a Jukes–Cantor model) with uniform root distribution. A polynomial F in the dummy variables $(t_g)_{g\in G^{\mathbf{L}}}$ will be such that $F((\mathbb{P}\{Y = g\})_{g\in\mathbb{G}^{\mathbf{L}}}) = 0$ for all choices of parameters in the model if and only if F is of the form*

$$F((t_g)) = \sum_{h\in\mathbf{H}} c_h h\left(\left(\sum_{g\in\mathbb{G}^{\mathbf{L}}} \langle g, \chi\rangle t_g\right)_{\chi\in\hat{\mathbb{G}}^{\mathbf{L}}}\right),$$

where only finitely many of the coefficients c_h are non-zero and $\Sigma_{h\in\mathbf{K}} c_h = 0$ for all $\mathbf{K} \in \mathscr{H}_2 \setminus \{\mathbf{H}_{2,\rho}\}$ (respectively, for all $\mathbf{K} \in \mathscr{H}_1 \setminus \{\mathbf{H}_{1,\rho}\}$).

The analogues of the discussion following Theorem 6.1 about how to generate equivalent multinomials are straightforward and are left to the reader.

Finally, we remark that the results for models which have an arbitrary root distribution are very similar to those above. The only difference is that the root is treated like a typical vertex in a Kimura three-parameter model. That is, instead of imposing the constraints $\mathbb{E}[\langle Z_\rho, \theta\rangle] = 0$, $\theta \in \{\phi, \psi, \phi\psi\}$, we use the fact (which follows a fortiori from Lemma 4.1) that the image of

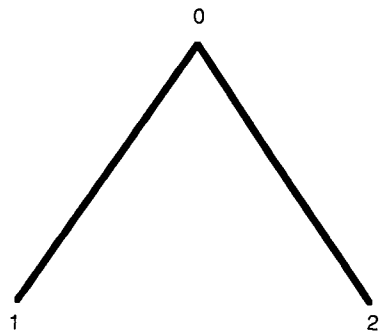

FIG. 4.

$(\mathbb{E}[\langle Z_\rho, \phi\rangle], \mathbb{E}[\langle Z_\rho, \psi\rangle], \mathbb{E}[\langle Z_\rho, \phi\psi\rangle])$ as the distribution of Z_ρ varies over all possible distributions has a nonempty interior. We once again leave the details to the reader. As an exercise, we invite the reader to show that there are no linear invariants for any Kimura three-parameter model with arbitrary root distribution.

7. Examples. In this section we present some explicit results for rooted trees with 2, 3 and 4 leaves, respectively, describing invariants under both the Kimura three-parameter and two-parameter models, with either a uniform or an arbitrary root distribution.

Two taxa. We begin with the simple two-leaf tree of Figure 4. Suppose (cf. Section 4) that Z_0, Z_1 and Z_2 are mutually independent $\mathbb{G}$-valued random variables (where, as before, $\mathbb{G} = \{A, G, C, T\}$) with distributions π, r_1 and r_2, respectively. Suppose that r_1 and r_2 are elements of R_3 with parameters $(\alpha_1, \beta_1, \gamma_1)$ and $(\alpha_2, \beta_2, \gamma_2)$ and a common value of t. Set $(Y_1, Y_2) = (Z_0 + Z_1, Z_0 + Z_2)$, so that we have a Kimura three-parameter model. The distribution $f = (f_{g_1, g_2})$ of (Y_1, Y_2) has Fourier transform

$$\hat{f}(\chi_1, \chi_2) = \hat{\pi}(\chi_1\chi_2)\hat{r}_1(\chi_1)\hat{r}_2(\chi_2),$$

where $\chi_1, \chi_2 \in \hat{\mathbb{G}}$. Expressions for $\hat{r}_1$ and $\hat{r}_2$ can be found in Section 4. When π is uniform, it follows that $\hat{f}(\chi_1, \chi_2) = 0$ unless $\chi_1 = \chi_2$, and the values of $\hat{f}(\chi, \chi)$ are just those of the Fourier transform of the convolution $r_1 * r_2$. It is easy to check that in this case only the parameters $t(\alpha_1 + \alpha_2)$, $t(\beta_1 + \beta_2)$ and $t(\gamma_1 + \gamma_2)$ are identifiable and so these, together with the requirement that the four distinct values of f_{g_1, g_2} must sum to $1/4$, account for all the degrees of freedom in this simple case. In what follows we will repeatedly use this informal counting of degrees of freedom, because in all cases we describe, the numbers match.

We turn now to the case where the initial distribution is arbitrary. The class of all such f is described by nine parameters: three for the root distribution,

and three each for the two edges. Given that $\Sigma f_{g_1,g_2} = 1$, we need to find a further six constraints and in this case they are all *nonlinear*. Following the construction outlined after the proof of Theorem 6.1, we can calculate the relevant 15×9 matrix of 0's and 1's, and it is easily checked that this matrix has row rank 9, that is, there are six linearly independent relationships defining invariants. There are many ways to describe six independent invariants, although, as explained in the proof, a minimal basis of the lattice could be constructed. We content ourselves here with presenting a typical cubic invariant:

$$\mathbb{E}\langle Y_1, \phi\rangle\mathbb{E}\langle Y_2, \psi\rangle\mathbb{E}[\langle Y_1, \psi\rangle\langle Y_2, \phi\rangle] = \mathbb{E}\langle Y_1, \psi\rangle\mathbb{E}\langle Y_2, \phi\rangle\mathbb{E}[\langle Y_1, \phi\rangle\langle Y_2, \psi\rangle]$$

and there are two others like this one.

Turning to the Kimura two-parameter model for this tree (still with an arbitrary root distribution), suppose that $\beta_1 = \gamma_1$ and $\beta_2 = \gamma_2$, that is, that r_1 and r_2 are elements of R_2. Recall that $\hat{r}_i(\phi) = \hat{r}_i(\theta)$, $i = 1, 2$, where we write θ for $\phi\psi \in \hat{\mathbb{G}}$. We can then see that

$$\hat{f}(\phi, \phi) = \hat{f}(\theta, \theta)$$

and

$$\hat{f}(\phi, \theta) = \hat{f}(\theta, \phi),$$

and these identities define two *linear* invariants. A simple counting argument would suggest that there are $16 - 1 - (3 + 2 \times 2) = 8$ independent invariants in all, leaving six *nonlinear* invariants to be found, and this can be verified by following the construction given in the proof of Theorem 6.2 and computing the row rank of the appropriate 15×7 matrix of 0's and 1's. In constructing a collection of six independent *nonlinear* invariants we find that some may be taken to be the same as those described earlier, whilst others result from the fact that $\hat{r}_i(\phi) = \hat{r}_i(\theta)$, $i = 1, 2$. This latter class includes the quadratic invariant

$$\mathbb{E}\langle Y_1, \theta\rangle\mathbb{E}\langle Y_2, \phi\rangle = \mathbb{E}\langle Y_1, \phi\rangle\mathbb{E}\langle Y_2, \theta\rangle.$$

Before closing this discussion of the invariants of the Kimura two-parameter model for the two-leaf tree, it is both of independent interest and convenient for later examples to relate our notation and results to those of Cavender (1989, 1991). Following Cavender, we let A (resp. G, C, T) double as the *function* on $\mathbb{G} = \{A, G, C, T\}$ which takes the value 1 on A (resp., G, C, T) and 0 elsewhere. We then see that $A - G = (1/2)(\phi + \theta)$ and $C - T = (1/2)(\phi - \theta)$; and, letting $\otimes$ denote the tensor product of functions on $\mathbb{G}$, we see that

$$\begin{aligned}(A - G) \otimes (C - T) &= \tfrac{1}{4}(\phi + \theta) \otimes (\phi - \theta) \\ &= \tfrac{1}{4}(\phi \otimes \phi - \phi \otimes \theta + \theta \otimes \phi - \theta \otimes \theta)\end{aligned}$$

and

$$\begin{aligned}(C - T) \otimes (A - G) &= \tfrac{1}{4}(\phi - \theta) \otimes (\phi + \theta) \\ &= \tfrac{1}{4}(\phi \otimes \phi - \theta \otimes \phi + \phi \otimes \theta - \theta \otimes \theta).\end{aligned}$$

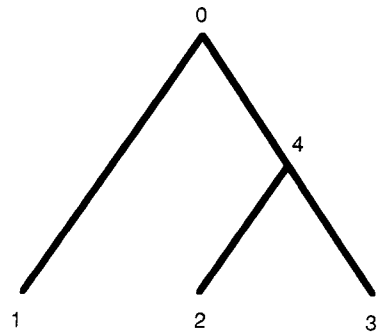

FIG. 5.

Letting f_{ij}, $i, j \in \{A, C, G, T\}$, denote the probability of observing base i in the first species and base j in the second species [cf. Cavender (1991)], we have

$$\begin{aligned} f_{AC} - f_{GC} - f_{AT} + f_{GT} &= \sum_{g_1, g_2} (A - G) \otimes (C - T)_{g_1, g_2} f_{g_1, g_2} \\ &= \sum_{g_1, g_2} \tfrac{1}{4}(\phi + \theta) \otimes (\phi - \theta)(g_1, g_2) f_{g_1, g_2} \\ &= \tfrac{1}{4}\big[\hat{f}(\phi, \phi) - \hat{f}(\phi, \theta) + \hat{f}(\theta, \phi) - \hat{f}(\theta, \theta)\big]. \end{aligned}$$

Similarly,

$$f_{CA} - f_{TA} - f_{CG} + f_{TG} = \tfrac{1}{4}\big[\hat{f}(\phi, \phi) - \hat{f}(\theta, \phi) + \hat{f}(\phi, \theta) - \hat{f}(\theta, \theta)\big].$$

Three taxa. We now discuss the rooted tree with 3 leaves, see Figure 5.

Again we will consider the Kimura three-parameter model first. Let $Z_0, \ldots, Z_4$ be mutually independent $\mathbb{G}$-valued random variables with Z_0 having distribution π, and Z_i having distribution $r_i \in R_3$, $1 \le i \le 4$. If we set $Y_1 = Z_0 + Z_1$, $Y_2 = Z_0 + Z_4 + Z_2$ and $Y_3 = Z_0 + Z_4 + Z_3$, then the distribution $f = (f_{g_1, g_2, g_3})$ of (Y_1, Y_2, Y_3) has Fourier transform

$$\hat{f}(\chi_1, \chi_2, \chi_3) = \hat{\pi}(\chi_1\chi_2\chi_3)\left[\prod_{i=1}^{3} \hat{r}_i(\chi_i)\right]\hat{r}_4(\chi_2\chi_3),$$

where $\chi_1, \chi_2, \chi_3 \in \hat{\mathbb{G}}$.

Suppose that π is uniform. Then the only nonzero values of the Fourier transform occur when $\chi_1\chi_2\chi_3 = 1$, that is, when $\chi_3 = \chi_1\chi_2$. There are 16 distinct probabilities with a single sum constraint, depending upon six parameters, and so we might expect to find nine independent invariants. Calculation of the rank of the appropriate matrix reveals this to be the case. All of these invariants are *nonlinear*, a typical one being

$$\begin{aligned} &\mathbb{E}[\langle Y_1, \theta\rangle\langle Y_2, \phi\rangle\langle Y_3, \psi\rangle]\mathbb{E}\langle Y_1 + Y_2, \psi\rangle\mathbb{E}\langle Y_1 + Y_3, \phi\rangle \\ &\quad = \mathbb{E}[\langle Y_1, \theta\rangle\langle Y_2, \psi\rangle\langle Y_3, \phi\rangle]\mathbb{E}\langle Y_1 + Y_2, \phi\rangle\mathbb{E}\langle Y_1 + Y_3, \psi\rangle. \end{aligned}$$

Now let us suppose π to be arbitrary. Then we have 64 probabilities with a single sum constraint, given in terms of 15 parameters: three for the root distribution and three for each of the four edges of the tree. We note at this point a difference between the situation where π is uniform and π is arbitrary. In the former, the edge between the root and vertex 4 does not count in the parametrisation; the tree is effectively unrooted or, equivalently, a star phylogeny. This is because the result of convolving the uniform distribution with any distribution on $\mathbb{G}$ is again the uniform distribution, and hence the distribution r_4 gets "lost," that is, its parameters are unidentifiable. When π is general, however, we do need to count the parameters of r_4. Doing so would suggest that there are $64 - 1 - 15 = 48$ independent invariants, and again a check of the appropriate matrix rank shows this to be the case. An example of an invariant in this case is

$$\mathbb{E}[\langle Y_1, \chi\rangle\langle Y_2 + Y_3, \chi'\rangle] = \mathbb{E}[\langle Y_1, \chi\rangle]\mathbb{E}[\langle Y_2 + Y_3, \chi'\rangle]$$

for all $\chi, \chi' \in \hat{\mathbb{G}}$, expressing the obvious independence of Y_1 and $Y_2 + Y_3$.

Turning to the Kimura two-parameter for this tree (still with an arbitrary root distribution), we impose the constraints $\beta_i = \gamma_i$, $i = 1, \ldots, 4$. We now find that there are 18 *linear* invariants, arising from equalities of the form $\hat{f}(\chi_1, \chi_2, \chi_3) = \hat{f}(\chi_1', \chi_2', \chi_3')$, where for $i = 1, 2, 3$, either $\chi_i = \chi_i'$, $(\chi_i, \chi_i') = (\phi, \theta)$ or $(\chi_i, \chi_i') = (\theta, \phi)$. In essentially the same notation as Cavender (1989, 1991) these invariants may be written as $X \otimes (A - G) \otimes (C - T)$, $X \otimes (C - T) \otimes (A - G)$, where $X \in \mathbb{G}$ is arbitrary, and two similar sets of pairs, with X occupying the second and third position in the triple. This identification is easily obtained using the relations given at the end of the discussion of the two taxa case.

Let us note here a difference between our analysis and a result stated by Cavender (1991). In that paper it is asserted that the space of linear invariants of the six-parameter Cavender–Lake model coincides with that of the Kimura two-parameter model. However, we can only find 12 linearly independent linear invariants for the Cavender–Lake model, compared with the 18 obtained above for the Kimura two-parameter model. Indeed it is not hard to check that the linear function with coefficients $(\phi + \theta) \otimes (\phi - \theta) \otimes \psi$ is simultaneously an invariant for the Kimura model, and an element of the "expected spectrum" for the Cavender–Lake model (i.e., a linear combination of joint probabilities under the model).

To confirm this last assertion, set

$$\begin{aligned} D &= \tfrac{1}{4}(\phi + \theta) \otimes (\phi - \theta) \otimes \psi \\ &= (A - G) \otimes (C - T) \otimes ((A + G) - (C + T)) = E - 2F, \end{aligned}$$

where

$$E = (A - G) \otimes (C + T) \otimes ((A + G) + (C + T))$$

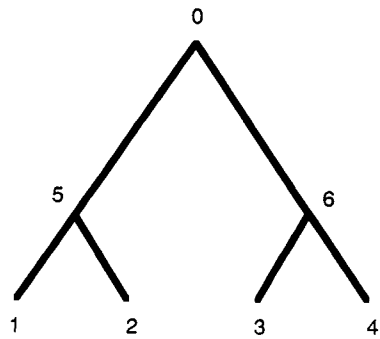

FIG. 6.

and

$$F = (A - G) \otimes C \otimes (C + T) + (A - G) \otimes T \otimes (A + C).$$

It is easy to check that D is an invariant for the two-parameter Kimura model, and we now show that E and F both belong to the expected spectrum of the Cavender–Lake model.

Firstly, consider $A \otimes (C + T) \otimes (A + G)$. Put $\pi = \delta_A$, the unit point mass at A. In the notation of Section 2, put $P^{(1)} = P^{(4)} = I$, the identity matrix. Recalling the basis $\{p_i\}$ given in Cavender (1989), put $P^{(2)} = p_3$ and $P^{(3)} = p_2$. We find that up to a scalar the result is $A \otimes (C + T) \otimes (A + C)$. Similarly, we can get $A \otimes (C + T) \otimes (C + T)$, $G \otimes (C + T) \otimes (A + G)$ and $G \otimes (C + T) \otimes (C + T)$, which implies that E is indeed in the expected spectrum of the Cavender–Lake model.

On the other hand, putting $\pi = \delta_A$, $P^{(1)} = P^{(2)} = I$, $P^{(4)} = p_3$ and $P^{(3)} = p_6$, shows that $A \otimes G \otimes (C + T) + A \otimes T \otimes (A + C)$ belongs to the expected spectrum. Similarly, replacing p_3 by p_2 shows that $G \otimes C \otimes (C + T) + G \otimes T \otimes (A + G)$ belongs to the expected spectrum. Thus F belongs to the expected spectrum and the assertion follows.

Four taxa. Our final example concerns the four-leaf tree given in Figure 6.

Yet again we consider the Kimura three-parameter model for this tree first. Let $(Z_i)_{i=0}^6$ be mutually independent $\mathbb{G}$-valued random variables, with Z_0 having distribution π and Z_i having distribution $r_i \in R_3$, $1 \leq i \leq 6$. Writing

$$Y_1 = Z_0 + Z_5 + Z_1,$$

$$Y_2 = Z_0 + Z_5 + Z_2,$$

$$Y_3 = Z_0 + Z_6 + Z_3,$$

$$Y_4 = Z_0 + Z_6 + Z_4,$$

we readily check that the Fourier transform of the distribution $f = (f_{g_1, g_2, g_3, g_4})$ of $(Y_i)_{i=1}^4$ factorises as follows:

$$\hat{f}(\chi_1, \chi_2, \chi_3, \chi_4) = \hat{\pi}(\chi_1\chi_2\chi_3\chi_4)\left[\prod_{i=1}^{4} \hat{r}_i(\chi_i)\right]\hat{r}_5(\chi_1\chi_2)\hat{r}_6(\chi_3\chi_4),$$

where $\chi_1, \chi_2\chi_3, \chi_4 \in \hat{\mathbb{G}}$.

When π is uniform, there are only 64 distinct probabilities in the 4^4 array f, corresponding to the 64 nonzero values of $\hat{f}$ which arise from quadruples $(\chi_1, \chi_2, \chi_3, \chi_4)$ such that $\chi_1\chi_2\chi_3\chi_4 = 1$. These quadruples are readily enumerated and it is easy to check that in this case there are 48 *nonlinear* independent invariants. We can write some of the invariants for this case in the following intuitively appealing forms:

(a) independence statements such as

$$\mathbb{E}[\langle Y_1 + Y_2, \chi\rangle\langle Y_3 + Y_4, \chi'\rangle] = \mathbb{E}[\langle Y_1 + Y_2, \chi\rangle]\mathbb{E}[\langle Y_3 + Y_4, \chi'\rangle]$$

for each of the *ordered* pair of nontrivial characters (χ, χ') (there are nine such equalities corresponding to the nine degrees of freedom in the usual chi-squared test of independence for a 4×4 contingency table);

(b) independence-like statements such as

$$\mathbb{E}\left\langle \sum_{i=1}^{4} Y_i, \chi \right\rangle \mathbb{E}\left\langle \sum_{i=1}^{4} Y_i, \chi' \right\rangle = \mathbb{E}[\langle Y_1 + Y_2, \chi\rangle\langle Y_3 + Y_4, \chi'\rangle] \times \mathbb{E}[\langle Y_1 + Y_2, \chi'\rangle\langle Y_3 + Y_4, \chi\rangle]$$

for each of the *unordered* pair of distinct nontrivial characters $\chi \neq \chi'$;

(c) equalities reminiscent of the determinantal identities on page 62 of Cavender and Felsenstein (1987) such as

$$\mathbb{E}[\langle Y_1 + Y_3, \chi\rangle\langle Y_2 + Y_4, \chi'\rangle]\mathbb{E}[\langle Y_1 + Y_3, \chi'\rangle\langle Y_2 + Y_4, \chi\rangle] = \mathbb{E}[\langle Y_1 + Y_4, \chi\rangle\langle Y_2 + Y_3, \chi'\rangle]\mathbb{E}[\langle Y_1 + Y_4, \chi'\rangle\langle Y_2 + Y_3, \chi\rangle]$$

for each of the *unordered* pairs of distinct characters $\chi \neq \chi'$ with not both χ and χ' trivial;

(d) cubic invariants obtained by considering three leaves at a time and using the invariants found in the three taxa case.

Once more the simple counting rules described earlier apply. There are $48 = 64 - 1 - 3 \times 5$ independent invariants, and we note as before that the uniform root distribution renders one of the distributions r_4 or r_5 superfluous.

Next we suppose π to be arbitrary. Then all six of the edge distributions contribute three parameters, as does the root distribution, and so there should be $256 - 1 - 3 - 6 \times 3 = 234$ independent invariants. The row rank of the appropriate 255×21 matrix is indeed 21 and so the counting rules continue to apply.

Finally, suppose that we have the Kimura two-parameter model (still with an arbitrary root distribution) obtained by fixing $\beta_i = \gamma_i$, $1 \leq i \leq 6$. There are 92 linearly independent *linear* invariants, in contrast with the 68 found by Cavender (1989) for the six-parameter Cavender–Lake model. These 92 linear invariants all arise from equalities of the form $\hat{f}(\chi_1, \ldots, \chi_4) = \hat{f}(\chi'_1, \ldots, \chi'_4)$ that occur when $(\chi'_1, \ldots, \chi'_4)$ is obtained from $(\chi_1, \ldots, \chi_4)$ by switching ϕ and θ, and such equalities are readily enumerated. Similarly, it is easy to check from the appropriate matrix that there are a total of $240 = 256 - 1 - 3 - 6 \times 2$ independent invariants, as expected.

Acknowledgments. We would like to thank Barbara Bowman, Arend Sidow and Allan Wilson for arousing our interest in this field, and for being generous with their time in explaining the area to us. A more immediate stimulus for this paper was the one-day workshop on invariants organised by Mike Waterman at the University of Southern California in July 1990. Thanks are due to him and James Lake, James Cavender and Joe Felsenstein, whose presentations and earlier work constitute the foundations on which we have built. Finally, we would like to thank Trang Nguyen for her assistance with the rank computations of Section 7.

REFERENCES

BARRY, D. and HARTIGAN, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Statist. Sci.* **2** 191–210.

CAVENDER, J. A. (1989). Mechanized derivation of linear invariants. *Molecular Biology and Evolution* **6** 301–316.

CAVENDER, J. A. (1991). Necessary conditions for the method of inferring phylogeny by linear invariants. *Math. Biosci.* **103** 69–75.

CAVENDER, J. A. and FELSENSTEIN, J. (1987). Invariants of phylogenies in a simple case with discrete states. *J. Classification* **4** 57–71.

COHN, H. (1980). *Advanced Number Theory*. Dover, New York.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.

DIACONIS, P. (1990). Patterned matrices. In *Matrix Theory and Applications—Proceedings of Symposia in Applied Mathematics* **40** (C. R. Johnson, ed.) 37–58. Amer. Math. Soc., Providence, RI.

DROLET, S. and SANKOFF, D. (1990). Quadratic tree invariants for multivalued characters. *Journal of Theoretical Biology* **144** 117–129.

FELSENSTEIN, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27** 401–410.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17** 368–376.

FELSENSTEIN, J. (1991). Counting phylogenetic invariants in some simple cases. *Journal of Theoretical Biology* **152** 357–376.

JUKES, T. H. and CANTOR, C. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (H. N. Munro, ed.) 21–132. Academic, New York.

KIMURA, M. (1980). A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16** 111–120.

KIMURA, M. (1981). Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. USA* **78** 454–458.

KIMURA, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press.

KÖRNER, T. W. (1988). *Fourier Analysis*. Cambridge Univ. Press.

LAKE, J. A. (1987). A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution* **4** 167–191.

NAVIDI, W., CHURCHILL, G. A. and VON HAESELER, A. (1992). Phylogenetic inference: Invariants and maximum likelihood. *Biometrics*. To appear.

NEYMAN, J. (1971). Molecular studies of evolution: A source of novel statistical problems. In *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.) 1–27. Academic, New York.

SANKOFF, D. (1990). Designer invariants for large phylogenies. *Molecular Biology and Evolution* **7** 255–269.

SWOFFORD, D. L. and OLSEN, G. J. (1990). Phylogeny reconstruction. In *Molecular Systematics* (D. M. Hillis and C. Moritz, eds.) 411–501. Sinauer, Sunderland, MA.

TAVARÉ, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on Mathematics in the Life Sciences* (R. Miura, ed.) 57–86. Amer. Math. Soc., Providence, RI.

DEPARTMENT OF STATISTICS
567 EVANS HALL
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720